DOCUMENT RESUME

ED 088 953 TM 003 521

AUTHOR Hanna, Gerald S.
TITLE Improving Reliability And Validity Of Multiple-Choice Tests With An Answer-Until-Correct Procedure.
PUB DATE Apr 74
NOTE 3p.; Paper presented at the joint session of the American Educational Research Association and the National Council on Measurement in Education (Chicago, Illinois, April 15-19, 1974)

EDRS PRICE MF-$0.75 HC-$1.50
DESCRIPTORS Feedback; *Multiple Choice Tests; Performance Factors; Response Style (Tests); Test Construction; Testing; *Test Reliability; *Test Validity
IDENTIFIERS Answer Until Correct; AUC

ABSTRACT

It was theorized that an answer-until-correct procedure, whereby an examinee marks responses to each multiple-choice question until feedback indicates that the correct answer has been marked, would yield scores of greater reliability and validity than conventional number-right procedure. Two papers and an application exercise for an undergraduate educational psychology class provided criterion measures with which validities of multiple-choice tests scores derived by each procedure were compared. Findings consistently favored the answer-until-correct method over number-right method in two reliability comparisons and in six validity comparisons. Importance and applications of findings are discussed. (Author)

ED 088953

TM 003 521

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IMPROVING RELIABILITY AND VALIDITY OF MULTIPLE-CHOICE TESTS WITH AN ANSWER-UNTIL-CORRECT PROCEDURE

Gerald S. Hanna
Kansas State University

The purpose of this study was to measure incremental reliability and validity resulting from a multiple-choice testing procedure whereby an examinee continues to select answers to each question until the correct answer is chosen.

In content areas assessed by best-answer type multiple-choice items, one is tempted to seek a finer discrimination among students on each question than that provided by the usual dichotomus scoring. Surely ability to answer a question correctly on the second trial implies more competence than ability to select the answer only after three or four attempts. Yet potential discrimination among examinees who fail to answer correctly on the first trial is sacrificed by conventional right-wrong scoring.

Since Pressey's (1926) early teaching-testing machine, many ingenious mechanical, electrical and chemical methods have been devised to provide better discrimination per item among students and/or to enable examinees to continue responding in a real-to-life fashion until feedback indicates that they are correct.

Gilman and Ferry (1972) recently reported a reliability increase from .79 to .93 resulting from the use of an Answer-Until-Correct (AUC) procedure. Their study raises two questions. First, is it reasonable to expect this much reliability increment to occur consistently? Probably not, but their success is luring. Second, is part of whatever reliability gain that can be expected a function of affective characteristics? It seems reasonable to speculate that the immediate feedback inherent in all varieties of AUC media may adversely affect the performance of some anxious examinees who happen to score poorly on the first few items. If it is true that internal consistency of cognitive achievement tests is raised as a result of consistent affective traits, then this increase in reliability is obtained only at the expense of construct validity; such reliability is not a virtue. These considerations emphasize the need for studies that investigate criterion-related validities of AUC devices against criteria for which this affective consideration is irrelevant.

Procedures

Thirty-eight undergraduate students in an educational psychology class

*Paper read at the annual meeting of the American Educational Research Association, Chicago, Illinois, April, 1974

provided data on (1) eleven 10-item multiple-choice quizzes, (2) a 50-item multiple-choice cumulative final examination, (3) two papers respectively dealing with behavioral objectives and with original examples of teaching for transfer of learning, and (4) an 82-item true-false interpretative exercise that relates human development, learning, and measurement to classroom applications.

In contrast to the usual directions and scoring methods for multiple-choice tests, the AUC method used with the above four-option, multiple-choice measures involves directing the examinee to indicate his answer to each question by erasing a carbon shield covering a feedback message. If the selected answer is correct on the first trial, he has completed the question; otherwise, he makes another response, etc., until the feedback message signifies that the correct answer has been selected. Scores were obtained by subtracting the sum of the total number of responses (erasures) made in finding the correct answer to every item from the total number of possible responses.

Also, an inferred (conventional) number-right (INR) score was obtained for each multiple-choice measure by counting the number of questions answered correctly on the first trial.

Criterion measures were such as to render method of scoring multiple-choice items irrelevant to their scores. Each of the two papers was subjectively evaluated on a numeric scale by the instructor. The true-false application exercise was administered without feedback and was scored objectively.

The odd-even reliability coefficient, corrected for full length, was computed for each multiple-choice measure scored by each method.

To provide validity measures, each quiz and the final examination scored by each method was correlated with each of the two papers and with the application exercise.

Findings and Conclusions

Table 1 summarizes the findings. The left-hand section displays means and standard deviations of the experimental variables. The middle section contains comparisons of the AUC and INR scores on odd-even reliabilities. The mean (computed by use of Fisher's z coefficients) reliability findings for the eleven quizzes are reported. In both reliability comparisons, the AUC procedure resulted in slightly higher internal consistency than the INR scoring method.

The right-hand side of Table 1 reports three criterion-related validity comparisons each for the mean of the quizzes and for the final examination. The comparisons shown in the six cells reveal slight to substantial superiority of the AUC method over the INR method. The validity increments for the mean of the quizzes are equivalent to what could be realized by lenthening the quizzes, scored by the INR method, by approximately 10 to 25 per cent. But the validity gains for the final examination could not have been achieved

by any amount of mere lengthening. Relevance to the criterion measures, as well as reliability, has been increased by the AUC procedures.

Collectively, these highly consistent, albeit statistically non-significant, findings suggest that the AUC procedure used in this study merits further study. With increasing availability of mechanical, chemical, and computerized testing devices, the economic viability of AUC methods is improving. If replications with varied test content, diverse examinees, and heterogeneous criterion measures consistently show superiority of AUC scoring over actual number right, as well as INR, scoring, then the method could provide a means of significantly enhancing the validity of multiple-choice testing. This improvement can be achieved with relative ease in situations (e.g., computer-managed instruction) wherein the immediate feedback provided by AUC procedures is desired in its own right for its instructional value.

Table 1

Reliability and Validity Comparisons

Multiple-Choice	Measures		Odd-Even	Validities		
Variable	M	S.D.	Reliability	Application Exercise	Paper I	Paper II
Mean Quiz AUC	26.1	2.7	.25	.33	.24	.25
Mean Quiz INR	7.4	1.5	.18	.31	.22	.23
Final AUC	125.8	11.0	.77	.42	.32	.19
Final INR	34.8	6.0	.76	.31	.22	.15

References

Gilman, D. A. and Ferry, P. Increasing Test Reliability Through Self-Scoring Procedures. Journal of Educational Measurement, 1972, 9, 205-207.

Pressey, S. L. A Simple Device Which Gives Tests and Scores and Teaches. School and Society, 1926, 23, 373-376.